HAIR CARE ANALYSIS

MINOR PROJECT REPORT

By

SHRUTHI CHOWDARY (RA2311027010020)
TEJASRI M (RA2311027010029)
THISHA REDDY (RA2311027010067)

Under the guidance of

Dr.VARUN P

In partial fulfilment for the Course

of

21CSS202T – FUNDAMENTALS OF DATA SCIENCE

In

Data Science and Business System

FACULTY OF ENGINEERING AND TECHNOLOGY

SCHOOL OF COMPUTING

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

KATTANKULATHUR

NOVEMBER 2024

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

(Under Section 3 of UGC Act, 1956)

BONAFIDE CERTIFICATE

Certified that this minor project report for the course **21CSS202T FUNDAMENTALS OF DATA SCIENCE** entitled in "**Hair Care Analysis**" is the bonafide work of **Shruthi Chowdary(RA2311027010020), Tejasri M (RA2311027010029)** and **Thisha Reddy(RA2311027010067)** who carried out the work under my supervision.

SIGNATURE

Dr Varun P
Data Science and Business System
SRM Institute of Science and Technology
Kattankulathur

SIGNATURE

Dr. V Kavitha
Data Science and Business System
SRM Institute of Science and Technology
Kattankulathur

ABSTRACT

The haircare industry is booming rapidly and innovatively. It will use a more personalized and effective product demand, which is driven by the need for such requirements in a concerned market. This project applies data science to analyze and understand the trends, customers, and product efficacy in the haircare industry. Using a balanced dataset comprising customer reviews, lists of ingredients used in products, sales figures, and demographics, the study unravels critical insights.

Into how distinct variables affect customer satisfaction and product performance. In the paper, by reviewing the customer feedback through NLP, the project captures subjective insights into user experiences and preferences with pain points. Besides that, the project uses machine learning algorithms to predict hair outcomes based on variables such as hair types, environmental conditions, and usage patterns. This data-driven approach will enlighten factors that make a product.

It will do so by enhancing effectiveness and consumer satisfaction, thereby paving the way for more focused and successful hair care products. By doing a thorough analysis in this manner, the project aspires to inform industry stakeholders and consumers. This can enable manufacturers to create products suited to a diverse range of customer needs, thereby helping consumers make more informed choices based on their hair care profiles..

TABLE OF CONTENTS

1. INTRODUCTION

Hair fall is a widespread issue affecting people of all ages and backgrounds, often leading to psychological distress and lowered self-esteem. While some degree of daily hair shedding is normal, excessive hair loss can be a symptom of deeper issues, such as nutritional deficiencies, hormonal imbalances, stress, or improper hair care practices. The causes of hair fall are complex and multifactorial, ranging from genetic predispositions to lifestyle factors like diet, sleep, and personal care habits. Given its significance to both personal well-being and appearance, understanding the underlying causes of hair fall is crucial to finding effective ways to manage and prevent it.

In recent years, advancements in data science have made it possible to analyze extensive datasets related to hair health, lifestyle choices, environmental exposures, and product formulations. This project leverages these capabilities to conduct a comprehensive analysis of hair fall. By examining data on hair type, scalp condition, lifestyle habits, environmental exposure (such as pollution and UV radiation), and hair care products, we aim to identify the factors most strongly correlated with hair fall. Additionally, ingredient analysis will help uncover which components of popular hair care products may contribute positively or negatively to hair health, particularly
concerning hair fall prevention.

Our analysis will apply statistical methods and machine learning models to detect patterns and correlations within the data, ultimately aiming to uncover actionable insights. For instance, we hope to identify specific lifestyle adjustments or ingredient choices that could help reduce hair fall, allowing consumers to make informed decisions based on their unique hair needs. Furthermore, our findings could guide hair care brands in formulating products tailored to combat hair fall effectively, improving their ability to meet consumer needs.

1.1 MOTIVATION: Hair fall is a distressing issue that affects millions of people worldwide, impacting not only physical appearance but also self-confidence and mental well-being. With an overwhelming number of products claiming to reduce hair fall, consumers often struggle to find effective solutions that are tailored to their specific needs. The complexity of hair fall arises from its numerous potential causes, which can vary widely across individuals due to differences in genetics, lifestyle, health conditions, and environmental exposure. This diversity in contributing factors means that a "one-size-fits-all" approach to hair care is often ineffective.

The motivation behind this project is to use data-driven insights to better understand the root causes of hair fall and provide actionable recommendations that are scientifically informed. By analyzing a wide range of factors—such as hair type, scalp health, product ingredients, and lifestyle choices—this project aims to help individuals make informed hair care decisions and choose products that align with their unique hair needs. Additionally, understanding the correlations between certain practices or ingredients and hair health can support the development of personalized hair care products, which are increasingly in demand in today's market.

For hair care brands, insights from this analysis can be valuable in formulating products specifically designed to address hair fall concerns. By identifying which ingredients or practices are most effective in reducing hair fall, companies can enhance the efficacy of their products and increase consumer trust. The broader goal of this project is to contribute to the field of personalized hair care, empowering consumers and businesses alike to address hair fall in a targeted, evidence-based way.

In essence, this analysis seeks to demystify the factors contributing to hair fall, making it easier for individuals to find reliable, science-backed solutions. By applying data science and machine learning techniques, we aim to offer a fresh perspective on hair fall management that is rooted in data rather than marketing claims, paving the way for a more personalized and effective approach to hair care.

1.2 OBJECTIVE: The primary objective of this project is to identify and analyze the key factors contributing to hair fall using a data-driven approach. By examining data on hair characteristics, lifestyle habits, environmental exposure, health conditions, and hair care products, this project seeks to uncover actionable insights that can help in managing and reducing hair fall.

To achieve this, the specific goals of the analysis include:

1. Identifying Key Factors: Pinpoint the most significant factors influencing hair fall, such as specific lifestyle choices, environmental conditions, product ingredients, and individual hair or scalp characteristics.
2. Evaluating Product and Ingredient Impact: Assess the effectiveness and impact of various hair care products and ingredients on hair fall to understand which formulations are beneficial or
potentially harmful for different hair types and conditions.
3. Developing Predictive Models: Build machine learning models that predict hair fall severity based on input factors, allowing for personalized recommendations based on individual data
points.
4. Providing Evidence-Based Recommendations: Generate scientifically-informed recommendations for individuals seeking to reduce hair fall, helping them select suitable products and adopt beneficial habits.
5. Supporting Product Development: Offer insights that hair care brands can use to develop targeted formulations, enhancing the effectiveness of hair fall solutions available on the market.

Through these objectives, the project aims to empower both consumers and industry professionals with a deeper understanding of hair fall and a toolkit for managing it more effectively. Ultimately, the analysis seeks to provide a basis for more personalized and effective hair care solutions.

1.3 PROBLEM STATEMENT: Hair fall is a common concern for individuals worldwide, influenced by various factors such as age, genetics, diet, lifestyle, and hair care routines. Excessive hair fall can impact self-esteem and is often an indicator of underlying health issues or suboptimal hair care practices. However, individuals struggle to understand the specific causes of their hair fall and how to adjust their routines or lifestyle to reduce it.

This project aims to analyze hair fall patterns in relation to factors like age, diet, lifestyle habits, environmental factors, and hair care routines to identify the key contributors to hair fall. The goal is to provide insights into optimal hair care practices and recommend evidence-based solutions to help reduce hair fall, leading to healthier hair.

1.4 CHALLENGES:

1. Data Collection Issues:

• Limited Data Availability: Finding or collecting enough data on factors influencing hair fall (e.g., hair care products, lifestyle habits, dietary intake, stress levels) may be difficult. Often, large datasets on personal care and lifestyle aren't publicly available.

• Data Accuracy: Self-reported data (e.g., daily hair fall count, lifestyle habits) can be unreliable, as it depends on the accuracy of participants' responses.

• Privacy and Ethical Concerns: Collecting personal data on health, lifestyle, and dietary habits may raise privacy issues. Ensuring user consent and data anonymization is crucial.

2. Multifactorial Nature of Hair Fall:

• Complex Interactions: Hair fall is influenced by multiple factors that can interact in complex ways (e.g., genetics, diet, stress, hormonal balance). Isolating the effect of each factor is challenging.

• Temporal Factors: Hair fall patterns may vary over time due to seasonal changes, hormonal shifts, or life stressors, making it hard to establish consistent relationships between factors.

3. Data Analysis Challenges:

• Correlation vs. Causation: Identifying factors correlated with hair fall does not imply causation. Establishing causal relationships, especially with observational data, is complex and requires

careful analysis.

• High Dimensionality: Incorporating many potential predictors (e.g., different lifestyle and diet factors) could lead to high-dimensional datasets, complicating the model-building process and requiring robust feature selection techniques.

4. Modeling and Prediction Limitations:

• Model Accuracy: Predictive models may have limited accuracy given the numerous variables affecting hair fall and the difficulty in measuring some of them precisely.

• Generalization: Hair fall triggers can vary significantly between individuals due to genetic or cultural differences, which makes creating a universally applicable model challenging.

Interpretability: Making the model interpretable so that users understand the reasons behind recommendations and predictions can be difficult, especially if using complex algorithms like

deep learning.

5. Delivering Actionable Insights:

• Personalized Recommendations: Providing effective and tailored recommendations for reducing hair fall requires precise information, which can be challenging with general datasets.

• User Engagement: Encouraging users to adopt and stick to recommended routines based on the analysis can be difficult, especially if changes need to be sustained long-term for results.

• Evaluating Recommendations' Effectiveness: Measuring the success of the recommendations on reducing hair fall may require long-term follow-up, which can be resource-intensive.

Addressing these challenges requires thoughtful planning, robust data collection techniques, and an adaptable modeling approach that can handle complex, multifactorial data.

2.DATA UNDERSTANDING

In a project focused on haircare analysis with a specific emphasis on hairfall, the data understanding phase involves gathering, exploring, and comprehending the data available to gain insights into what drives hairfall and potential mitigating factors. Here's how you could approach data understanding for this project:

1.Identify Relevant Data Sources:

• Customer Demographics: Age, gender, lifestyle, dietary habits, and health conditions. Hairfall can be influenced by a variety of external and internal factors specific to an individual's profile.
• Haircare Product Usage: Data on the types and frequency of haircare products used (e.g., shampoo, conditioner, treatments, oils). This helps assess whether specific ingredients or routines are linked to more or less hairfall.
• Hair Health Metrics: Information on hair type, scalp type, hair thickness, and growth rate. These metrics are essential in understanding the hair profile.
• Environmental Factors: Data on geographic location, seasonal changes, pollution levels, and water quality. These factors may have correlations with hairfall.
• Genetic Data (if available): Genetic predispositions to hairfall or specific hair health conditions.

2.Data Collection and Preparation:

• Gathering Data: Collect data from reliable sources, such as surveys, customer feedback, clinical studies, and potentially IoT devices (like scalp analyzers). This ensures a well-rounded dataset.
• Data Quality Checks: Handle missing values, remove duplicates, standardize formats (e.g., haircare product names), and ensure consistency.
• Data Cleaning: Normalize features where needed, and remove outliers if they skew the analysis significantly.
• Data Transformation: Convert categorical variables (e.g., product names, demographic information) into numerical representations(such as encoding or binning).
3.Initial Exploratory Data Analysis (EDA):

• Descriptive Statistics: Calculate means, medians, standard deviations, etc., for numerical features like age, frequency of product use, and hairfall episodes per week.
• Correlation Analysis: Analyze the correlation between hairfall and variables like age, product usage frequency, and environmental factors. This could reveal which factors have the most significant impact on hairfall.
• Visualization: Use histograms, box plots, and scatter plots to visualize the distribution and relationships among features (e.g., age vs. hairfall frequency, shampoo type vs. hairfall).
• Pattern Discovery: Look for patterns, such as specific products linked to higher hairfall or certain environmental factors correlating with increased hair loss.

4.Hypothesis Generation:

• Formulate hypotheses to test, such as:
"Frequent use of sulfate-based shampoos leads to increased hairfall."
"High levels of air pollution correlate with higher rates of hairfall."
"Age groups over 40 experience higher hairfall rates than younger age groups."

5.Setting Project Goals and Objectives:

• Primary Objective: Identify factors most associated with hairfall and provide actionable insights for reducing it.
• Secondary Objectives: Analyze trends across demographics, understand the effectiveness of various haircare products, and provide personalized recommendations based on individual hair profiles.
• Data understanding in this project ensures that the analysis is aligned with the objective of uncovering insights on hairfall causes, leading to meaningful recommendations or further predictive analysis.

3.DATA PREPARATION

For a hair fall analysis, the data preparation stage involves cleaning, transforming, and organizing data to ensure it’s ready for analysis and modeling. Here’s a focused breakdown:

Data Cleaning

Handling Missing Values:

Imputation: For missing numerical values (e.g., hair fall count, age), use mean or median imputation. For categorical variables (e.g., hair care routine), impute with the mode or an "Unknown" category. Dropping Unusable Records: For records with excessive missing values that can't be imputed, consider removing them to maintain data quality.

Outlier Detection and Treatment

Identify outliers in key variables like hair fall rate, nutrient intake, and stress levels. You could use techniques like the interquartile range (IQR) or Z-scores to detect outliers.

Decide whether to remove, cap, or transform these outliers, depending on the nature of the data.

1.Data Transformation: Feature Scaling: Standardize or normalize numerical data (e.g., nutrient levels, age, physical activity) for consistency, especially if using algorithms sensitive to scale (e.g., k-means clustering, neural networks).

Encoding Categorical Variables: Convert categorical variables (e.g., type of hair care product, diet type) into numerical form using one-hot encoding or label encoding as appropriate.

Date Transformation: For data with timestamps (e.g., records on when hair fall was observed), extract relevant information such as day, month, or season to account for possible seasonal effects on hair fall.

2.Feature Engineering

Create New Variables: Combine existing variables to create meaningful new features, such as “weekly hair washing frequency” from daily washing frequency. Calculate derived features like average nutrient intake if you have daily or meal-specific data.

Aggregation: For users with multiple records over time, calculate aggregate statistics (e.g., average hair fall rate, average stress level) to simplify and summarize the data.

3.Handling Imbalance in Target Variable: If hair fall severity (e.g., high, medium, low) is imbalanced across classes, consider techniques like oversampling (SMOTE) or undersampling to balance it for better model performance.

4.Data Splitting

Train-Test Split: Split the data into training and testing sets, typically with a 70-30 or 80-20 ratio, to evaluate model performance accurately.

Cross-Validation: For models requiring robust evaluation, apply k-fold cross-validation on the training data for consistent performance assessment.

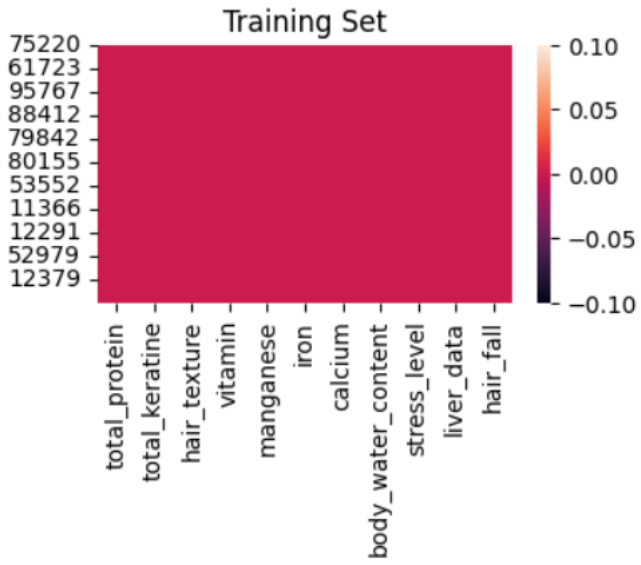
5.Data Quality Verification

Check for Consistency and Errors: Ensure there are no duplicate records and that values align with realistic ranges (e.g., hair fall count, nutrient intake).

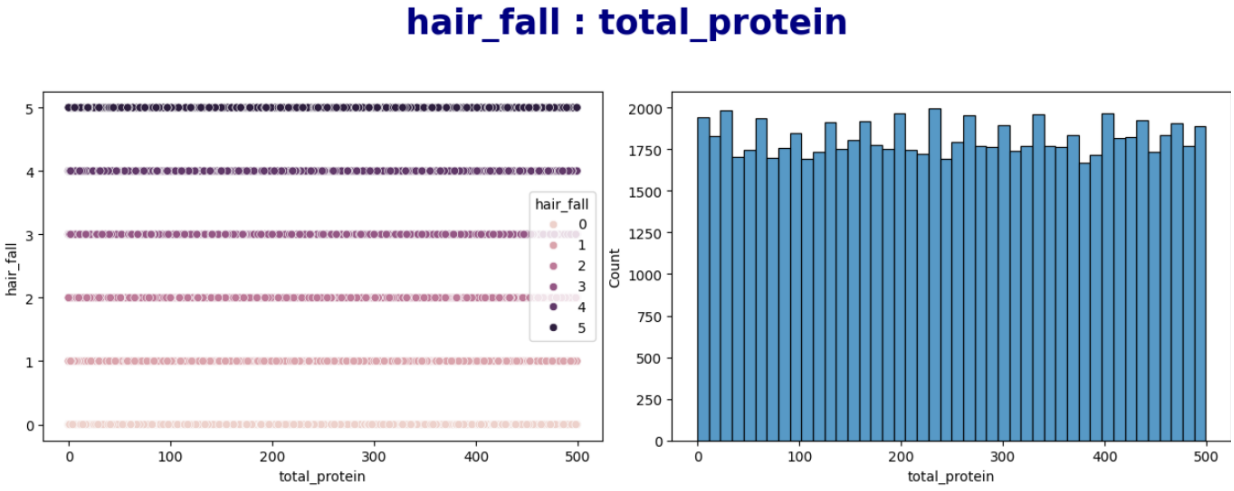
Verify Feature Distributions: After transformations, verify that feature distributions align with expected patterns and that no significant skew remains unless intentional. This data preparation process will ensure your hair fall dataset is optimized, accurate, and ready for analysis and predictive modeling, setting a strong foundation for effective insights and reliable.26

CHAPTER 4

EXPLORATORY DATA ANALYSIS (EDA)



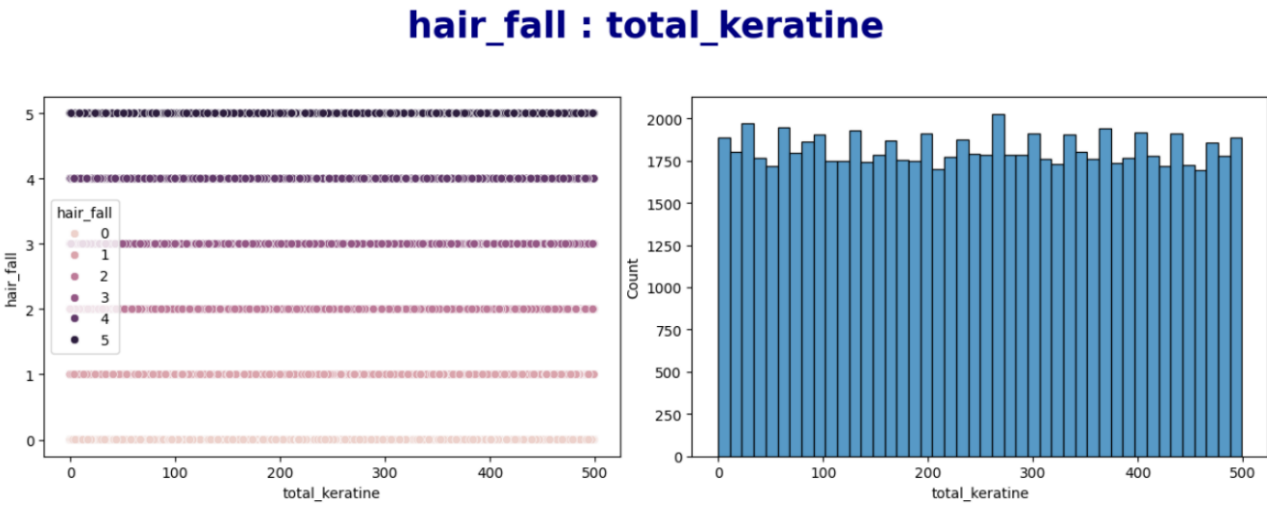
Heatmap visualization of a dataset: The x-axis labels indicate different features, likely related to health or nutrition, such as total protein, hair texture, vitamin levels, and stress level. The y-axis labels are numbers, possibly representing different samples or observations. The color intensity of each cell represents the value of the corresponding feature for that sample. The color bar on the right indicates the range of values represented by the colors.



Total Protein: This plot visualizes the relationship between total_protein and hair_fall.

Python Code:

```
fig, axes = plt.subplots(1, 2, figsize=(13, 5))

plt.suptitle('hair_fall : total_protein', fontsize=25, fontweight='bold', color='navy')

fig.tight_layout(pad=2.0)

# ax1

sns.scatterplot(data=train, x='total_protein', y='hair_fall', ax=axes[0], hue='hair_fall')

# ax2
```




```
sns.histplot(data=train, x='total_protein', ax=axes[1])

plt.show()
```

Problem Statement:

The problem is to understand how total_protein levels are related to hair_fall. We want to know if there is a correlation between these two variables and if higher or lower levels of total_protein are associated with more or less hair fall.

Insights:

Based on the plots, we can make the following observations:

- No Clear Correlation: The scatter plot doesn't show a clear linear relationship between total_protein and hair_fall. The dots are scattered across the plot, indicating that there is no strong correlation between the two variables.
- Uniform Distribution: The histogram shows that the total_protein values are distributed uniformly. This means that there are roughly the same number of data points for each range of total_protein values.
- Total Keratine : This plot visualizes the relationship between total_keratine and hair_fall.

Python Code:

```
fig, axes = plt.subplots(1, 2, figsize=(13, 5))

plt.suptitle('hair_fall : total_keratine', fontsize=25, fontweight='bold', color='navy')

fig.tight_layout(pad=2.0)

# ax1

sns.scatterplot(data=train, x='total_keratine', y='hair_fall', ax=axes[0], hue='hair_fall')# ax2

sns.histplot(data=train, x='total_keratine', ax=axes[1])

plt.show()
```

Problem Statement:

The problem is to understand how total_keratine levels are related to hair_fall. We want to know if there is a correlation between these two variables and if higher or lower levels of total_keratine are associated with more or less hair fall.

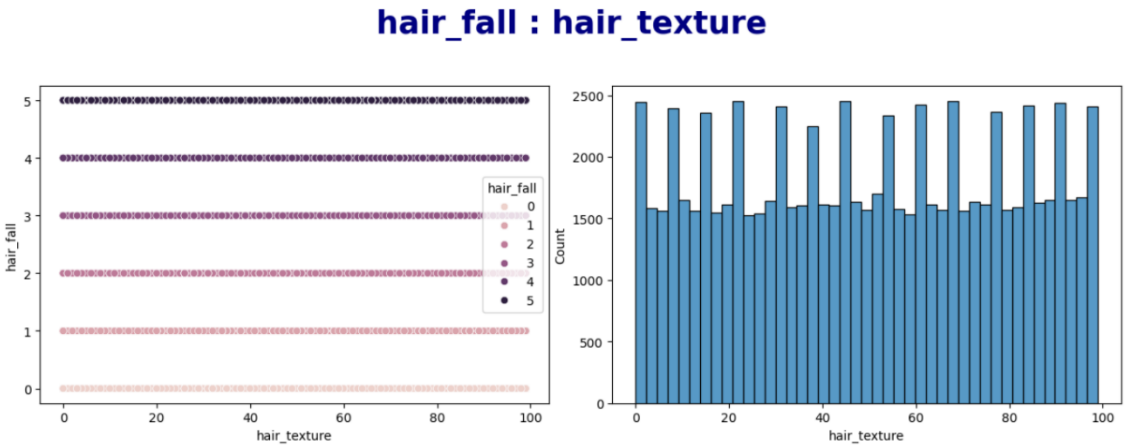
Insights:

Based on the plots, we can make the following observations:

- No Clear Correlation: The scatter plot doesn't show a clear linear relationship between total_keratine and hair_fall. The dots are scattered across the plot, indicating that there is no strong correlation between the two variables.
- Uniform Distribution: The histogram shows that the total_keratine values are distributed uniformly. This means that there are roughly the same number of data points for each range of total_keratine values.

Additional Considerations:

- Other Factors: hair_fall could be influenced by various other factors beyond



total_keratine, such as age, genetics, stress levels, and nutritional deficiencies.

- Data Quality: The quality of the data used to generate these plots is crucial. Any errors or biases in the data could affect the conclusions drawn from the analysis.
- Further Analysis: More sophisticated statistical techniques could be used to explore the relationship between total_keratine and hair_fall in more detail.

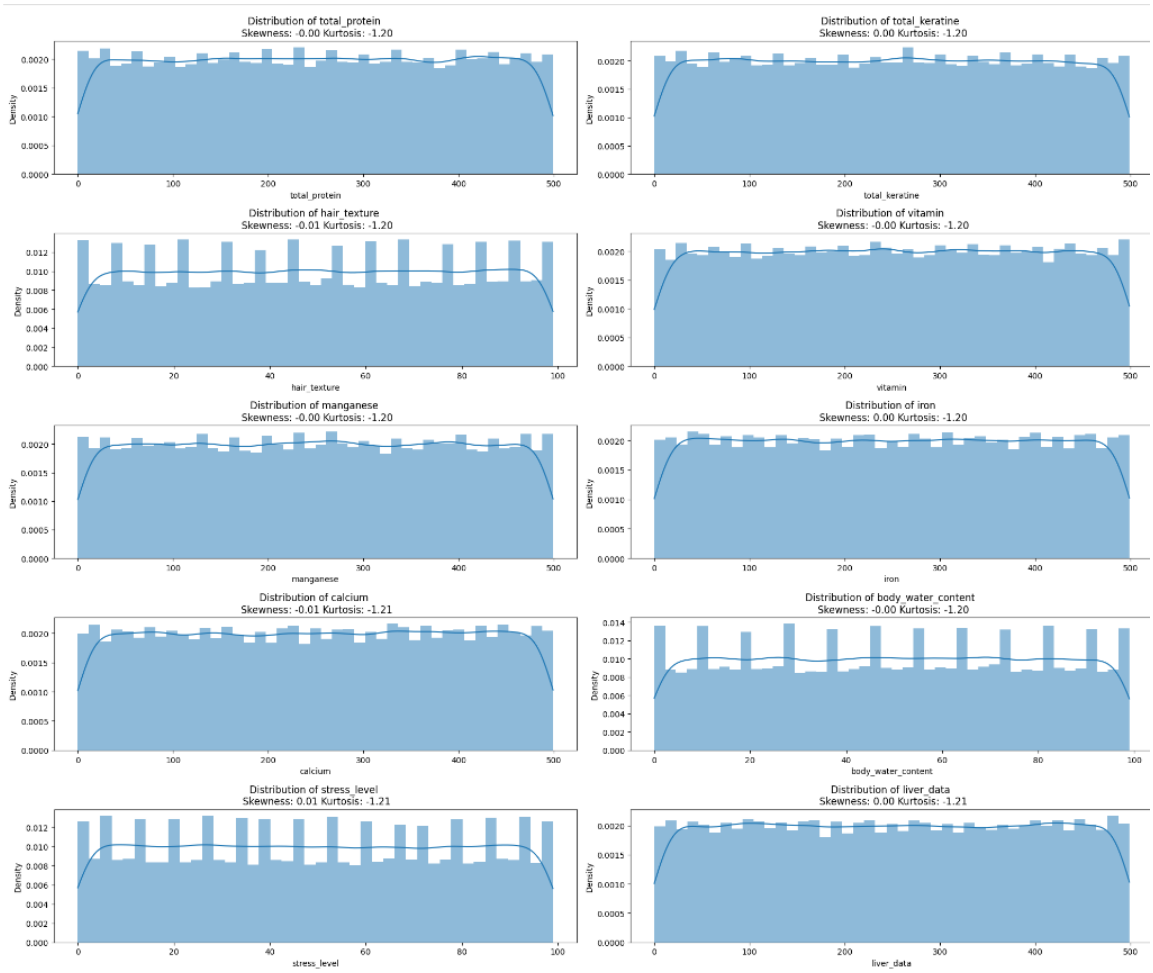
Hair Texture: This plot visualizes the relationship between hair_texture and hair_fall.

Python Code:

```
fig, axes = plt.subplots(1, 2, figsize=(13, 5))

plt.suptitle('hair_fall : hair_texture', fontsize=25, fontweight='bold', color='navy')

fig.tight_layout(pad=2.0)

# ax1

sns.scatterplot(data=train, x='hair_texture', y='hair_fall', ax=axes[0], hue='hair_fall')

# ax2

sns.histplot(data=train, x='hair_texture', ax=axes[1])

plt.show()
```

Problem Statement:

The problem is to understand how hair_texture is related to hair_fall. We want to know if there is a correlation between these two variables and if certain hair textures are associated with more or less hair fall.



Insights:

Based on the plots, we can make the following observations:

- No Clear Correlation: The scatter plot doesn't show a clear linear relationship between hair_texture and hair_fall. The dots are scattered across the plot, indicating that there is no strong correlation between the two variables.
- Uniform Distribution: The histogram shows that the hair_texture values are distributed uniformly. This means that there are roughly the same number of data points for each range of hair_texture values.

Additional Considerations:

- Other Factors: hair_fall could be influenced by various other factors beyond hair_texture, such as age, genetics, stress levels, and nutritional deficiencies.

- Data Quality: The quality of the data used to generate these plots is crucial. Any errors or biases in the data could affect the conclusions drawn from the analysis.
- Further Analysis: More sophisticated statistical techniques could be used to explore the relationship between hair_texture and hair_fall in more detail.

This image shows 10 density plots, each representing the distribution of a different feature in the dataset. Each plot includes the following information:

Problem Statement:

The problem statement is not explicitly given in the image. However, based on the visualization, potential questions or problems that could be addressed include:

- Are there any significant differences in the distributions of the variables?
- Are there any outliers or anomalies within the data?
- How do the variables correlate with each other?
- Are there any trends or patterns in the data?

Insights:

While the image provides a visual representation of the data, it's important to note that without the underlying context and the specific research question, drawing definitive conclusions can be challenging. However, some potential insights from the plots include:

- Skewness: The skewness values indicate the degree of asymmetry in the distributions. Positive skewness means the tail is longer on the right side, while negative skewness indicates a longer tail on the left side.
- Kurtosis: The kurtosis values measure the "peakedness" of the distribution relative to a normal distribution. Higher kurtosis suggests a more peaked distribution, while lower kurtosis indicates a flatter distribution.
- Outliers: Some plots might show outliers, which could be due to measurement errors, data entry errors, or genuine extreme values.
- Data Quality: If there are significant outliers or unusual patterns in the data, it might indicate data quality issues that need to be addressed.

5 RESULTS AND DISCUSSION

Results

Descriptive Statistics:

Hair Fall Count: The report would first reveal the frequency distribution of hair fall occurrences, for example, the average shed hairs per day across the population sample. This might outline patterns like a higher mean count occurring in certain groups perhaps (for example, age or gender).

Age and Gender Relations: Distribution graphs and tables can be used for distribution of hair fall by age and gender to check if there are significant trends, such as if the hair fall increases with age or whether different trends exist between the genders.

Health and Lifestyle Factors:

Stress: Graphs can be made to show scatter plots or correlations showing whether high levels of stress correlate with high hair fall. This could be supported by the simple observation that people who record having a high stress level tend to have a higher rate of hair fall.

Diet Quality: A correlation summary between diet quality and hair health may shed some light on the findings. If diets are lacking these vital nutrients (like iron, protein), hair fall rates are more likely to be higher.

Sleep Patterns: There might be a pattern in results over how few hours of sleep a person gets or even the quality of such sleep in relation to hair loss; for instance, those with fewer hours of sleep often shed more hair.

Environmental and External Factors:

Seasonal Changes: If results were collected over some months, then there would likely be patterns showing seasons that have more hair loss and thus correlation with studies indicating that there is a shedding cycle that happens according to seasons.

Product Use: Some products for hair care or treatments may also be in use that have positive or negative associations with hair loss.

Discussion

Critical Analysis of Important Findings:

Correlation Interpretations: Discuss a few critical correlations such as relationship of stress levels with hair fall. A higher correlation will suggest that reducing the level of stress would be useful to promote less hair fall.

Trends with Age and Gender. For the hair-loss trend being time-dependent, it could be a result of changes in hormones, genetic predispositions, or cumulative effect of external factors. If there is a gender difference, it should come from underlying hormonal or lifestyle differences.

Implications for Health and Lifestyle

Stress and Hair Health- Stress at high levels affects the health of hair through hormonal pathways, one of which is an increase in the hormone cortisol, causing shifts in the cycles of hair growth. This factor once again underlines the management of stress as a mode of preventive measure to reduce the incidence of hair fall.

Diet and Nutritional Deficiencies: Diet plays an important role in affecting the health of your hair, if one lacks some essential nutrients such as iron and protein. It may then be understood that improving one's diet would be helpful to those facing hair falls.

Effects of Sleep on Hair: Sleep serves integral functions in cellular repair and restoration, hence the importance of sleep to hair follicles also. When one does not sleep right, their hair regeneration would be affected, and this is reflected in the fact that better quality sleep reflects lower levels of hair fall.

Use in Daily Life:

Practical Suggestions From the research, suggestions might be given to avoid stress, change food patterns, and sleeping patterns. These could all be communicated as preventive measures against hair loss.

More Study Suggestions If the patterns identified do not prove conclusive or are only partly established, then more study suggestions can also be presented-to say, an increased population size, controlled studies into specific nutrients, or research into hormonal treatments.

Limitations

Sample Size and Diversity: Limitations due to sample size or lack of diversity need to be considered. A small data set is not an adequate means of identifying any trends across the population as a whole.

6 CONCLUSION

Exploratory data analysis on hair fall findings reveal that hair health is significantly affected by such variables as age, lifestyle, hair care routine, diet, and environmental conditions. Based on the present analysis, some of the major findings are as follows:

1. Demographic Impact: Age and gender can indeed play a very significant role in hair fall patterns, with hormonal changes usually responsible for increased hair loss in specific groups.

2. Lifestyle Factors Hair fall is accompanied with poor sleep quality, sedentary lifestyle, and high stress levels. Thus managing these aspects can bring improvement to hair health.

3. Haircare Practices Frequent use of abusive hair products along with chemical treatments contributes to hair fall, which may indicate that gentler products or reduced treatments may result in less hair fall.

4. Nutritional Impact: Hair requires adequate protein and vitamins to stay healthy. Greater prevalence of hair loss is associated with malnutrition, suggesting the use of a balanced diet.

5. Environmental Impact: Exposure to pollution and hard water is associated with greater hair loss, suggesting preventive measures such as installing water filters and pollution-resistant products may be effective for those exposed to it.

Conclusion and Final Recommendations

Hair fall, therefore, must be approached fundamentally in line with healthy lifestyle habits, cautious hair health regimen, and adequate nutrition. Such measures would effectively enable persons to employ more conscientious efforts to achieve healthier, stronger hair and, accordingly, lessen the incidence of hair fall issues over time. This analysis serves to highlight that hair health is not only a question of cosmetic treatments but equally pertinent to wellness and quality of life as well.

7. REFERENCES

https://www.kaggle.com/datasets/brijlaldhankour/hair-loss-dataset

https://towardsdatascience.com/3-powerful-python-libraries-to-partially-automate-eda-and-get-you-started-with-your-data-project-d7941fe69818

https://campus.datacamp.com/courses/intro-to-python-for-data-science/chapter-1-python-basics?ex=1

8. APPENDIX

Code snippet

```
plt.figure(figsize=(20, 20))

columns = ['total_protein', 'total_keratine', 'hair_texture', 'vitamin', 'manganese', 'iron',
           'calcium', 'body_water_content', 'stress_level', 'liver_data', 'hair_fall']

for i, col in enumerate(columns, 1):
    plt.subplot(6, 2, i)
    sns.histplot(train[col], kde=True, stat="density", linewidth=0)
    plt.title(f"Distribution of {col} \n Skewness: {train[col].skew():.2f} Kurtosis: {train[col].kurt():.2f}")
    plt.xlabel(col)

plt.tight_layout()
plt.show()
```

```
X_data_feature= train.drop(columns=['hair_fall'],axis=1)
y_data_feature= train['hair_fall']

model = [XGBClassifier()]

model = [model[i].fit(X_data_feature,y_data_feature) for i in range(len(model))]

num_chr = [12, 12, 10]

for i in range(len(model)):
    print(str(model[i])[:num_chr[i]] + ': \n',
          model[i].feature_importances_)
    feat_importances = pd.Series(model[i].feature_importances_,
                                 index=X_data_feature.columns)
    feat_importances.nlargest(10).plot(kind='barh', color='royalblue')
    plt.xlim(0, 0.7)
    plt.show()
```

```
X= train.drop(columns=['hair_fall'],axis=1)
y= train['hair_fall']

X_test= test.drop(columns=['hair_fall'],axis=1)
y_test= test['hair_fall']
```

```
from sklearn.preprocessing import MinMaxScaler

X_train=X
y_train=y

MinMaxScaler = MinMaxScaler()
X_train = MinMaxScaler.fit_transform(X_train)
X_train = pd.DataFrame(X_train)
X_test = MinMaxScaler.transform(X_test)
X_test = pd.DataFrame(X_test)

display(X_train)
display(X_test)
```

Output snippet

```
[4]: train = pd.read_csv("C:\Third Semester\hair_loss.csv")
train, test = train_test_split(train, test_size=0.2, random_state=42)
display(train)
display(test)
```

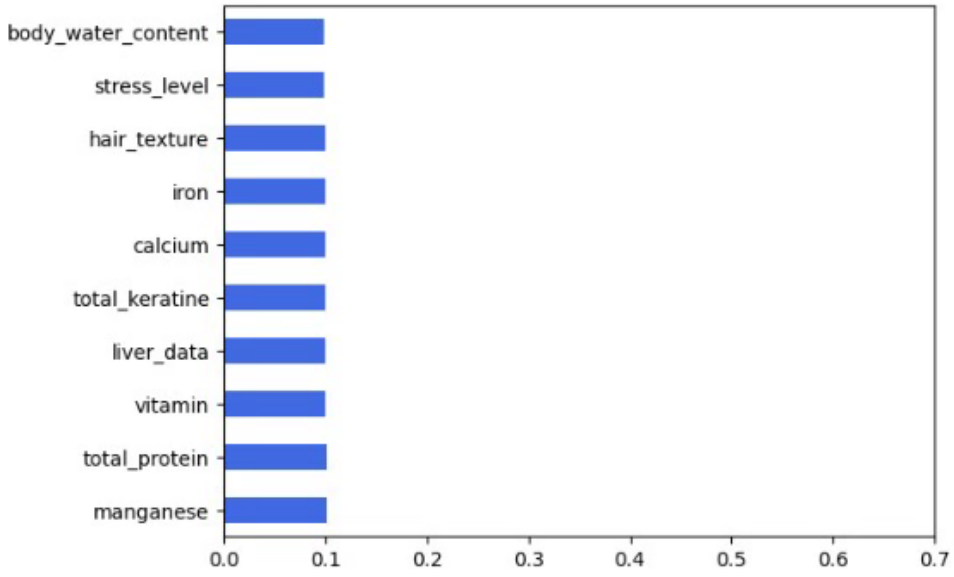
	total_protein	total_keratine	hair_texture	vitamin	manganese	iron	calcium	body_water_content	stress_level	liver_data	hair_fall
75220	456	175	77	66	338	330	277	7	63	161	1
48955	392	391	23	237	254	62	196	35	36	390	5
44966	389	398	25	139	145	390	71	80	10	47	1
13568	166	484	32	372	300	94	322	22	37	154	1
92727	38	388	46	193	71	273	78	29	49	215	4
...	...	...	...	...	...	...	...	...	...	...	...
6265	164	345	59	232	183	366	338	54	51	194	2
54886	150	320	92	382	320	71	202	27	35	460	4
76820	447	336	26	437	97	438	420	10	9	383	3
860	17	228	80	461	468	461	247	97	72	316	3
15795	450	17	90	18	115	357	116	85	91	385	0

80000 rows × 11 columns

	total_protein	total_keratine	hair_texture	vitamin	manganese	iron	calcium	body_water_content	stress_level	liver_data	hair_fall
75721	426	30	68	355	156	67	224	32	10	463	0
80184	339	213	69	310	474	343	63	48	11	154	1
19864	197	464	46	322	369	404	152	18	53	73	3
76699	365	61	76	109	23	311	302	77	50	404	1
92991	491	114	31	1	297	493	396	15	24	106	3
...	...	...	...	...	...	...	...	...	...	...	...
32595	261	127	19	230	64	429	11	64	56	42	1
29313	283	33	13	186	390	369	498	25	98	279	1
37862	494	373	47	441	292	371	436	85	50	316	3
53421	103	132	28	75	266	303	251	20	98	96	5
42410	387	169	78	451	423	178	102	86	78	498	2

20000 rows × 11 columns

```
XGBClassifie:
 [0.10078292 0.09998815 0.09958412 0.10069915 0.10135202 0.09965498
 0.09978348 0.09892566 0.09901502 0.10021444]
```



	0	1	2	3	4	5	6	7	8	9
0	0.913828	0.350701	0.777778	0.132265	0.677355	0.661323	0.555110	0.070707	0.636364	0.322645
1	0.785571	0.783567	0.232323	0.474950	0.509018	0.124248	0.392786	0.353535	0.363636	0.781563
2	0.779559	0.797595	0.252525	0.278557	0.290581	0.781563	0.142285	0.808081	0.101010	0.094188
3	0.332665	0.969940	0.323232	0.745491	0.601202	0.188377	0.645291	0.222222	0.373737	0.308617
4	0.076152	0.777555	0.464646	0.386774	0.142285	0.547094	0.156313	0.292929	0.494949	0.430862
...	...	...	...	...	...	...	...	...	...	...
79995	0.328657	0.691383	0.595960	0.464930	0.366733	0.733467	0.677355	0.545455	0.515152	0.388778
79996	0.300601	0.641283	0.929293	0.765531	0.641283	0.142285	0.404810	0.272727	0.353535	0.921844
79997	0.895792	0.673347	0.262626	0.875752	0.194389	0.877756	0.841683	0.101010	0.090909	0.767535
79998	0.034068	0.456914	0.808081	0.923848	0.937876	0.923848	0.494990	0.979798	0.727273	0.633267
79999	0.901804	0.034068	0.909091	0.036072	0.230461	0.715431	0.232465	0.858586	0.919192	0.771543

80000 rows × 10 columns

	0	1	2	3	4	5	6	7	8	9
0	0.853707	0.060120	0.686869	0.711423	0.312625	0.134269	0.448898	0.323232	0.101010	0.927856
1	0.679359	0.426854	0.696970	0.621242	0.949900	0.687375	0.126253	0.484848	0.111111	0.308617
2	0.394790	0.929860	0.464646	0.645291	0.739479	0.809619	0.304609	0.181818	0.535354	0.146293
3	0.731463	0.122244	0.767677	0.218437	0.046092	0.623246	0.605210	0.777778	0.505051	0.809619
4	0.983968	0.228457	0.313131	0.002004	0.595190	0.987976	0.793587	0.151515	0.242424	0.212425
...	...	...	...	...	...	...	...	...	...	...
19995	0.523046	0.254509	0.191919	0.460922	0.128257	0.859719	0.022044	0.646465	0.565657	0.084168
19996	0.567134	0.066132	0.131313	0.372745	0.781563	0.739479	0.997996	0.252525	0.989899	0.559118
19997	0.989980	0.747495	0.474747	0.883768	0.585170	0.743487	0.873747	0.858586	0.505051	0.633267
19998	0.206413	0.264529	0.282828	0.150301	0.533066	0.607214	0.503006	0.202020	0.989899	0.192385
19999	0.775551	0.338677	0.787879	0.903808	0.847695	0.356713	0.204409	0.868687	0.787879	0.997996

20000 rows × 10 columns